ChinaXiv合作期刊
知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum
E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

基于主题模型和时间序列分析的新兴主题识别与特征关联研究

李雅倩[1,2] 孙玉玲[1,2] 赵婉雨[1]
1. 中国科学院文献情报中心 北京 100090
2. 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100049

摘要：[目的/意义]开展新兴主题识别研究，科学有效地发掘其特征关联规律，可以更好地服务于现实需求，发挥科技情报研究对学科发展的创新支撑作用。[方法/过程]从新兴主题特征定义出发，结合新兴主题研究与科学影响评价的相关理论与实践，利用自然语言处理、全局主成分分析和时间序列分析方法建立新兴主题识别的方法框架，量化主题的一致性、新颖性、影响力和增长性等特征，结合趋势预测完成对新兴主题的提取、分析和识别。在新兴主题识别的基础上，深度挖掘目标领域新兴主题发展的规律，利用格兰杰因果检验和协整分析，对其特征关联效应进行长期均衡检验和因果关系推断，分析影响新兴主题发展的长期关联因素及其作用关系。[结果/结论]提出一套新兴主题识别及其关联特征分析的方法。为证实该方法的可行性和有效性，选取湿地领域开展实证研究，结合主题识别与特征关联效应分析，刻画该领域主题科学影响的动态发展路径，从关联特征视角出发提出新兴主题的建设思考。

关键词：趋势预测 新兴主题识别 特征关联效应 协整分析 面板数据分析
分类号：G20

引用格式：李雅倩, 孙玉玲, 赵婉雨. 基于主题模型和时间序列分析的新兴主题识别与特征关联研究 [J/OL]. 知识管理论坛, 2022, 7(3): 229-247[引用日期]. http://www.kmf.ac.cn/p/289/.

1 引言

随着科学研究第四范式的兴起，数据驱动下的科学研究正从知识层下沉到数据层，制定科技发展计划和相关政策需要紧随科研动态。文献作为知识流动的重要载体，是识别学科主题的重要数据来源。面对海量文本数据，如何科学有效地从中发掘新兴研究主题，是科研管理者和科研人员研究方向布局和调整的重要参考依据[1]。同时，学科主题发展具有“惯性”与“相关性/延续性”，即学科主题时间序列变化发展具有延续性并且是相互联系的，一定时期内存在可预测的发展变化规律。新兴主题的识别和

作者简介：李雅倩，硕士研究生；孙玉玲，副研究员，硕士生导师，通信作者，E-mail：sunyl@mail.las.ac.cn；赵婉雨，馆员。

收稿日期：2021-12-01 发表日期：2022-05-11 本文责任编辑：刘远颖

趋势预测，有助于科研工作者了解研究动态，有利于基金资助组织和决策者优化创新资源分配，进一步促进有发展潜力的研究方向的发展。

与新兴主题相似的概念较多，诸如热点主题、前沿主题和颠覆式主题等，并由此演化出一般创新主题、新兴前沿主题和科学前沿等概念，在研究和应用中容易出现概念边界的模糊不清 [2]。H. Xu 等计量“新兴主题”相关概念族群的研究热度和发展趋势，指出族群概念间存在差异和交叉，相比于前沿主题和颠覆式主题等，学者们对于新兴主题的研究兴趣增长更快 [3]。新兴主题相关概念的差别主要体现在时间维度和创新维度，热点主题、新兴主题和前沿主题在时间指向上，分别表征过去、现在和未来的重要研究主题，其创新程度随时间发展逐渐增强，预测难度也逐渐增大 [4]。

在新兴主题的识别方法上，学者们主要利用共词分析 [5]、引文分析 [6] 和文本挖掘分析 [7] 等相关技术方法，从科学文献中提取和识别新兴主题。近几年，针对新兴主题特征的讨论越来越多，大多数学者聚焦于文献的外部历史性特征，如文本主题的历史演化和引用情况等，而对于未来发展趋势的考虑较少 [8]。王山 [2] 认为，新兴主题代表着研究领域的未来趋势，对其趋势的分析解读尤为重要。随着相关研究热度持续增长，识别方法也日益多元化和科学化，但是在新兴研究主题的明确概念定义与拟定的可操作性指标之间仍然缺乏良好的联系 [9]。因此，如何挖掘新兴主题与特征之间的关联关系，采取有效的特征方案，通过构建科学严谨的预测模型和使用合适的分析方法进而提取长期关联变量，可以为新兴主题识别提供一些参考。笔者从新兴主题的全面特征出发，利用自然语言处理和时间序列趋势模型方法，综合分析文本和特征数据，进行新兴主题识别及特征关联分析。

2 相关研究与主要进展

新兴主题识别可以及时跟踪科学发展动态，尽早捕捉相关领域未来的发展契机和可能的变化趋势 [10]。梳理新兴主题概念和识别方法研究，相关进展大致可分为 3 类：面向新兴主题定义及其特征的研究讨论、面向新兴主题识别方法的融合创新和面向主题趋势分析的综合运用。

2.1 新兴主题概念及特征

1965 年，D. J. De Solla Price[11] 开创性地定义了研究前沿，认为这是一种处于正在成长阶段的新颖性研究。新兴主题起源于对研究前沿的关注 [12]，在新兴主题完整概念提出前，主题特征识别处在萌芽阶段，主要表现为采用多指标计量进行特征识别，如 R. L. Ohniwa 等 [12] 认为主题词增长性和丰富性是表征新兴主题的重要信息；Y. N. Tu 等 [13] 认为新颖性和研究热度是新兴主题最显著的特征。

2015 年，D. Rotolo 等 [14] 对新兴技术主题提出了全面的特征定义，考虑到技术和科学的差异性，Q. Wang[15] 对新兴主题进行了定义，即新兴主题是具有新颖性和一定连贯性、能产生较大科学影响力且发展速度相对较快的主题，其 4 个主要的特征分别为：新颖性、增长性、一致连贯性和科学影响力。伴随完整概念的提出，新兴主题特征分析迈入新阶段。H. Xu 等 [3] 提出针对新兴主题的多维科学计量指标评价方案，其中，新颖性和增长被认为是新兴主题的最重要指标，这两个指标被视为阈值指标，在确保新颖性和增长的前提下，考虑了对社会和经济以及对社区网络结构的显著影响的潜力。新兴主题的研究价值来源于其未来的增长潜力或科学影响潜力 [2]。S. Xu 等关注新兴主题的未来趋势，通过分析主题特征走势并预见新兴的研究主题 [9]。新兴主题特征定义被提出后，新兴主题研究取得了新的进展，一方面有关学者不断探索新的定义以及新的识别方法，另一方面一些学者致力于开发一系列的识别指标 [3]。

2.2 新兴主题识别方法

经过不断发展和创新，新兴主题识别方法经历由单一方法到机器学习、文本挖掘等多元化方法的融合。H. Small[16] 首次提出利用共引识

知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

别新兴主题，C. Chen[17] 将引文与词法分析结合，联合引文分析和爆破检测识别新兴主题。文本挖掘可以细粒度地挖掘大规模语料库中的文本关系特征 [18]，M. Blei 等先后提出的主题模型 [19] 和动态影响模型 [20] 等，可根据概率突发和关联规则识别领域新兴主题 [21]，获得了较为广泛的使用。

近年来，学者们在文本挖掘方法的基础上，探索基于新兴主题特征的多维特征的识别方法。李静等根据内外部文本特征构建新兴主题综合识别公式 [22]；白敬毅等 [23] 将主题新颖性、增长性、影响力等特征指标依次赋权叠加，利用多维尺度绘制主题分布矩阵识别新兴主题；S. Xu 等 [9] 利用动态影响模型提取主题结构及增长性和影响力等指标，使用多任务最小二乘支持向量机区分不同主题的特征表现等。如能融合多维特征构建综合识别方案，将有助于更好地实现新兴主题识别。

2.3 主题趋势预测

在新兴主题识别的研究中，越来越多的学者关注到主题的趋势特征。A. Kontostathis 等 [24] 观测词频趋势判定新兴主题；C. Lee 等 [25] 使用多层神经网络来捕获一定时段内关联指标间的非线性关系，开发了两个衡量主题趋势的定量指标。针对主题时间序列数据，岳丽欣等利用 ARIMA（Autoregressive Integrated Moving Average model）模型分别预测了热点主题 [26] 和主要研究主题 [27] 的未来趋势；刘自强等 [28] 运用 ARDL 模型度量主题趋势和扩散滞后效应，可见，时间序列分析方法已经取得了一些应用。

目前新兴主题概念及特征已经较为清晰，虽然不少学者考虑到趋势因素，但主要为了对研究现状进行分析解读，而对未来趋势变化的预测稍显不足。在新兴主题识别中，普遍采用综合识别公式等方法，一定程度上压缩了主题特征，对主题特征的动态变化过程有待进一步研究。笔者在 Q. Wang 等 [15] 提出的新兴主题基本定义的基础上，加入时间序列分析对主题趋势进行预测，作为潜在高成长性特征，结合全局主成分分析，从全领域视角分析各个主题的特征水平，系统地构建影响力和增长性的综合评价指标体系，结合时间序列方法进一步分析主题成分的动态特征，以对相关领域主题的特征表现情况及其深层次的关系进行剖析。

3 新兴主题识别方法框架

笔者提出的新兴主题识别与分析框架主要分为 4 个部分（见图 1）。针对文本数据，利用 LDA 主题识别生成主题时间序列，结合 ARIMA 模型和全局主成分量化主题特征，构建新兴主题识别方案。在新兴主题识别的基础上，综合采用面板协整分析和格兰杰因果推断，挖掘观测变量间的长期关系和关联效应，分析新兴主题及其特征的长期关联关系。

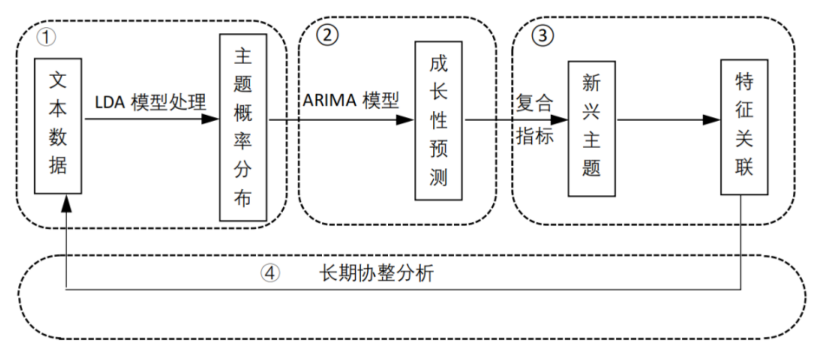


图 1 研究技术路线

3.1 主题识别和数据提取

笔者主要使用 python 语言进行摘要文本数据分词、清洗和词形还原，在与作者关键词、文章关键词合并去重后，通过 LDA 主题模型获



知识管理论坛
2022年 第3期(总第39期)

取主题概率分布。选择主题数量为1-175个的模型，经一致性比较和人工核验，确定最优主题数量。根据主题模型导出分布结果，计算主题各维度数据。

3.2 基于多维指标的新兴主题识别体系构建

目标领域主题的发展存在多种多样的外在体现，笔者从新兴主题定义出发，确定了基于新兴主题特征的量化指标识别体系，即在一致连贯性和新颖性指数基础上，采用ARIMA模型对主题未来成长潜力的预测结果，联合影响力和增长性特征时序立体表进行创新的全局主成分分析，刻画主题发展的动态特征与综合表现，综合各维度特征完成新兴主题的识别，如图2所示：

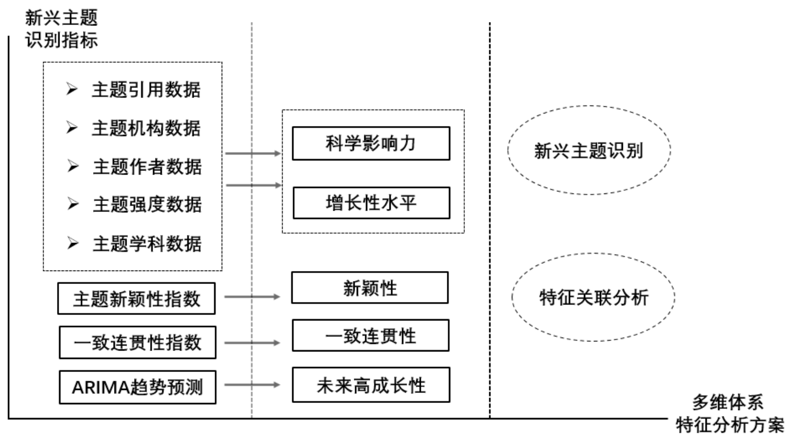


图2 新兴主题多维特征识别体系

3.2.1 未来高成长性

未来高成长性是指主题在未来具有良好的发展潜力。本文主要采用ARIMA模型，从主题强度数据出发预测其未来趋势。ARIMA(p，d，q)模型包括AR过程、MA过程和差分整合过程，内含3个主要参数分别为：p为自回归项数，d为平稳差分阶数，q为滑动平均项数[31]。ARIMA模型可以表示为：

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \quad \text{公式（1）}$$

在公式（1）中，L是滞后算子，d ∈整数，d>0。

3.2.2 新颖性

新颖性的度量是新颖主题识别的关键部分。Y. N. Tu等[13]利用发文时间测算新颖性指数。白敬毅等[23]增加主题生命周期理论，定义新颖性，如公式（2）所示：

$$NI=1/(t-FY+1) \quad \text{公式（2）}$$

其中，t为主题生命周期，FY为主题初次出现年份。考虑到湿地领域存在长生命周期主题，为保证区分度采用主题平均年龄，计算公式为：

$$NI=1/(t-AY+1) \quad \text{公式（3）}$$

在公式（3）中，AY为加权主题平均年份，$AY=\sum_{m=1}^{M^{(t)}} p_{k,m}^{(t)} Y$。

3.2.3 一致连贯性

一致连贯性是指主题已经出现一段时间，且拥有持续稳定发展的趋势。Q. Wang等[15]将主题连贯性定义为主题链接的松散程度，通过领域内引文数与发文数量之比（一致性指数）来测量，并将阈值取为1。S. Xu[9]认为连贯性取决于主题提取方法是否可以确保提取的主题足够连贯。白如江等[29]认为，可以通过时间切片，在连续时间区间达到设定标准的主题，即为满足连贯性要求。本文综合采用相邻时间划片以及一致性指数计算方法，度量一致连贯性特征。

3.2.4 科学影响力及增长性

科学影响力评估范式包括数量、质量和效果论，涉及研究成果产生条件、呈现载体和传

ChinaXiv合作期刊
知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

播 3 个维度，以及研究强度、研究绩效、研究支撑能力、研究机构多样化程度和研究成果的传播能力等指标 [30]。对于新兴主题科学影响力的分析存在单一和多种指标的情况，如 Q. Wang 等 [15] 利用主题被引次数计算科学影响力；G. González-Alcaide 等 [31] 分析研究主题领域文献发布情况、合作者特征（作者、机构和国家）和施引文献等影响传播特征，评估研究成果在研究领域的影响。本研究立足于科学影响典型评价范式，选择引文数量、作者数量、机构规模以及学科丰富性等作为科学影响力的综合观测指标。

主题增长是一个增量的概念，可以从多个角度来衡量，如 Q. Wang 等 [15] 采用发文数量增长，H. Guo 等 [32] 分析突发关键词、作者数量以及跨学科性等特征变化。结合相关研究内容，兼顾指标的实用性和可获得情况，笔者围绕主题影响力和增长性两方面的内涵，主要选择能够体现主题使用热度、研究参与热度、研究关注热度和研究增长情况方面的指标，在通过全局主成分分析的适宜性检验后，最终选取主题强度、主题引用、主题作者、主题机构和学科数量 5 项主题影响力评价指标，以及主题增长评价的 5 个增量指标，包括主题强度增长率，主题文献引用增长率、作者增长率、机构增长率和学科增长率，通过时序全局主成分分析得到评价综合影响力和综合增长性的两个主成分。针对主题特征的综合分析涉及多维面板数据的处理。全局主成分分析在处理此类数据上可以保留主题的动态特征，更具稳健性和代表性 [33]。

具体指标计算方法如下：

（1）主题引用指标。笔者认为，主题引用指标（Topic Citation，TCI）可以反映主题所拥有的关注度和传播热度，计算公式如下：

$$TCI_{k,t}=\sum_{m=1}^{M^{(t)}} c_{k,m}^{(t)} \qquad \text{公式（4）}$$

其中，$c_{k,m}^{(t)}$代表主题 k 在 t 年的第 m 个文档上的被引频次，按照文档年份进行同一主题下的频次累积加总即为主题引用指标。t 代表年份，m 为文章篇数，k 为主题个数。

（2）主题作者数量指标。主题作者数量（Topic Author index，TAT）指标测量的是某特定年份下，参与某个主题研究的学者的规模，也能在一定程度上反映出主题的热度。计算公式如下：

$$TAT_{k,t}=\sum_{m=1}^{M^{(t)}} au_{k,m}^{(t)} \qquad \text{公式（5）}$$

其中，$au_{k,m}^{(t)}$代表主题 k 在 t 年的第 m 个文档上的所有作者数量，其增长一方面来源于发文数量的增加，另一方面来源于参与研究人员数量的增加。

（3）主题学科数量指标。主题学科数量（Topic Category index，TCG）可以反映出主题学科跨度和学科交叉程度，笔者在增长性等指标设计上增加学科交差性等指标。计算公式如下：

$$TCG_{k,t}=\sum_{m=1}^{M^{(t)}} ct_{k,m}^{(t)} \qquad \text{公式（6）}$$

其中，$ct_{k,m}^{(t)}$代表主题 k 在 t 年的第 m 个科技文献的学科分类数量，笔者通过学科数量频次提取，按照文档年份累积加总得到主题学科数量指标。

（4）主题机构数量指标。主题机构数量（Topic Institution index，TIS）可以反映出学术机构对该领域的参与度，这也能反映出机构的研究方向选择和支持力度。该指标越大说明科研机构中在该主题下的布局越多。计算公式如下：

$$TIS_{k,t}=\sum_{m=1}^{M^{(t)}} inst_{k,m}^{(t)} \qquad \text{公式（7）}$$

其中，$inst_{k,m}^{(t)}$代表主题 k 在 t 年的第 m 个文档上的机构覆盖数量。

（5）主题强度指标。主题强度（Topic Indensity，TI）反映科技文献数据的研究热度，由各个文档的主题及其权重分布计算得到。B. Chen 等 [34] 研究发现，研究主题 k 在 t 时间的主题强度 $TI_{k,t}$，计算公式为：

$$TI_{k,t}=\sum_{m=1}^{M^{(t)}} p_{k,m}^{(t)} \qquad \text{公式（8）}$$

其中，$p_{k,m}^{(t)}$代表主题 k 在 t 年的第 m 个文档上的主题概率，该指标越大说明研究价值和研究意义越大。$TI_{k,t}$ 代表主题 k 在 t 年的第 m 个文档上的主题强度。

（6）增长性的度量。增长性体现在引文增长、作者增长、机构规模扩大 TI 以及不同学科

的汇集等方面，其度量方式为相邻时间数据的变化。通过计算，得到 TI-G、TIS-G、TCI-G、TCG-G 和 TAT-G，分别表征相应特征的增长。

以主题强度增长为例，度量公式为：

$$\text{Growth}(k,t) = (\text{TI}_{k,t}-\text{TI}_{k,(t-1)}) \quad \text{公式（9）}$$

增长性的计算方案如图 3 所示：

$p_{1,k}^{(1)}$	···	$p_{1,k}^{(t)}$	···	$p_{1,k}^{(T)}$
···		···		···
$p_{m,k}^{(1)}$	···	$p_{m,k}^{(t)}$	···	$p_{m,k}^{(T)}$
···		···		···
$p_{M^{(1)},k}^{(1)}$	···	$p_{M^{(1)},k}^{(t)}$	···	$p_{M^{(1)},k}^{(T)}$

TI(k,1)	...	TI(k,t)	...	TI(k,T)
$\sum_{m=1}^{M^{(1)}} p_{m,k}^{(1)}$	...	$\sum_{m=1}^{M^{(t)}} p_{m,k}^{(t)}$	...	$\sum_{m=1}^{M^{(T)}} p_{m,k}^{(T)}$

Growth(k,2)	···	Growth(k,t)	···	Growth(k,T)
TI(k,2)-TI(k,1)	···	TI(k,t)-TI(k,t-1)	···	TI(k,T)-TI(k,T-1)

图 3 主题强度增长计算演示

3.3 主题特征关联分析

为了深度挖掘目标领域新兴主题发展的内在发展规律，本研究采用主题特征关联分析方法。2003 年诺贝尔获奖者 C. W. J. Granger 提出基于“预测”的协整分析与格兰杰因果检验方法，对变量间的长期作用关系提出统计学上的检验，判断变量间的因果关联关系 [35]。对于包含截面个体特征与时间维度变化特征的面板数据，C. W. Kao 等提出面板协整检验的方法 [36]，E. I. Dumitrescu 和 C. Hurlin[37] 拓展了面板数据格兰杰因果关系的检验方法，从而可以更好地分析自变量与被解释变量的关联效用作用机制。针对新兴主题特征关联效应的分析，本研究主要采用上述方法。

4 新兴主题识别实证分析

4.1 数据来源

笔者利用“湿地”领域研究论文数据开展实证分析，在 Web of Science 平台核心合集的 SCIE 数据库（SCI-Expanded）和 SSCI 数据库（Social Sciences Citation Index）中进行检索。梳理湿地的不同类型和表达，并利用相关关键词设计检索策略，将标题、摘要、作者关键字和关键字作为识别字段，以 TI=((wetlands or wetland or “wet land” or “wet lands” or marsh or swamp* or peatland* or “peat land*” or bog or bogs or mire or mires or fen or fens or everglade* or mangrove*)) not TS=(“swamp crayfish*” or “marsh sandpiper” or “marsh mallow” or “marsh harbour”) 作为检索式进行主题检索，检索年代范围限制在 2000 年 1 月 1 日到 2020 年 12 月 31 日，检索时间为 2020 年 9 月，选取文献类型为“article”和“review”的文章，共计检索得到湿地领域相关文献 24 449 篇。论文年度分布情况见图 4，态势发展良好，增量稳步上升。

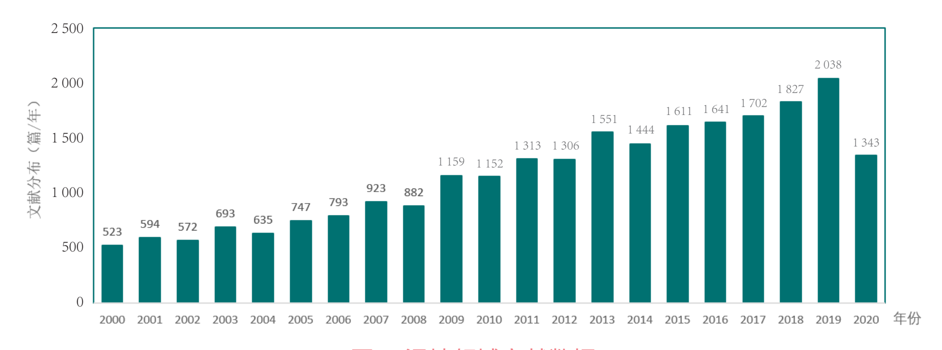


图 4 湿地领域文献数据

ChinaXiv合作期刊
知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

4.2 主题探测

笔者利用python进行主题识别，选择主题数量为1-175个的模型，综合比较困惑度（perplexity）和一致性的表现。其中，困惑度是利用概率计算某个主题模型在测试集上的表现，其值越低，则说明这个主题模型越好。困惑度分析结果表明，困惑度指标区分度不显著。C_v、U_mass、C_npmi和C_uci coherence均为一致性指标，衡量主题内词语之间是否为相互支撑关系，在一致性指标结果中，主题数目为26个时最优，见图5。

通过分词和主题模型等自然语言处理后导出主题—关键词分布，得到湿地领域的26个研究主题（见表1）。结合人工判读并翻译，湿地领域包括人工湿地再生、湿地生态监测、环境气候变化响应、湿地污染成分分析、湿地生物多样性保护、湿地气体排放通量模型与监测、退化湿地系统恢复、湿地循环系统分析、区域湿地管理、湿地恢复标准技术和湿地生态防护等主题。

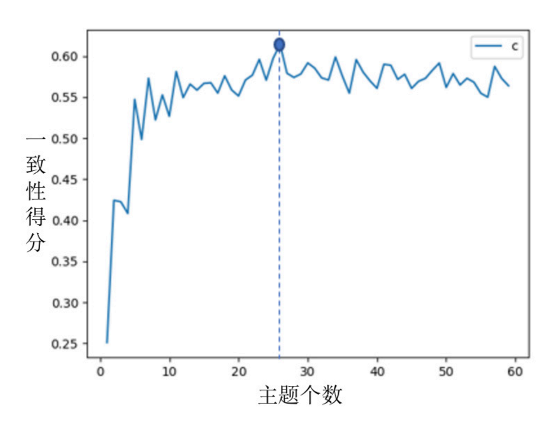


图5 主题—一致性可视化

表1 湿地领域研究主题—关键词列表

主编号	主题归纳	英文关键词
Topic1	人工湿地再生	removal|wetland|nitrogen|constructedwetlands|performance|system|rightsreserved|phosphorus
Topic2	湿地生态监测	wetland|vegetation|species|water|rightsreserved|diversity|soil|dynamics|site|restoration
Topic3	环境气候变化响应	coastalwetland|climatechange|marsh|sealevelrise|saltmarsh|partinaalterniflora|inundation|erosion
Topic4	湿地污染成分分析	pb|heavymetal|zn|cu|cd|contamination|cr|ni|mn|bioaccumulation|surfacesediments|fe|ca|bioavailability|hg
Topic5	湿地生物多样性保护	conservation|area|biodiversity|region|management|china|climate|landscape|agriculture|bird
Topic6	湿地气体排放通量模型与监测	co2|ch4|vulnerability|sequestration|carbondioxide|n2o|limited|microbialbiomass|tissues|feature
Topic7	退化湿地系统恢复	ecosystemservices|phytoremediation|accounting|wetlandecosystem|metrics|upland|ecosystemfunctions
Topic8	湿地循环系统分析	chemicaloxygendemandcod|adsorption|effluents|tidalwetlands|bod|hydrodynamics|phylogeneticanalysis
Topic9	区域湿地管理	hydraulicretentiontime|surfacewater|compounds|chemicaloxygendemand|archaea|sustainabledevelopment|cr
Topic10	湿地恢复的标准和技术	carbonsequestration|ch4fluxes|moisture|modelresults|river-basin|forestedwetland|slr|decompositionrates
Topic11	湿地生态防护	bacterialcommunity|hydraulicretentiontimehrt|hrt|microbialdiversity|aquaticenvironment|industrialwastewater
Topic12	湿地微生物群落研究	biofilm|synthesis|verticalflow|functionalgene|nest|subset|importantecosystems|mesocosm|protein|strain

续表 1

Topic13	湿地微生物基因研究	genes|waves|strains|dissolvedoxygen|bacterialdiversity|parasites|enzymeactivities|nacl|acid-minedrainage
Topic14	湿地生物种群趋势预测分析	landsat|timeseries|taxonomy|ammonianitrogen|paranariver|coleoptera|murray-darlingbasin|prescribedfire
Topic15	湿地生态补偿	northeastchina|dom|urban|mammals|wetlandprotection|ammonia-oxidizingbacteria|wetlandbirds|southernbrazil
Topic16	湿地分类与定量勘查研究	yellowriverdelta|remotesensingdata|liver|buffalo|swampeel|sewage-treatment|linearregression|landscapepattern|c/nratio
Topic17	湿地系统发生分析	bacterial|velocity|disease|co2fluxes|power|changeclimate|enzymeactivity|contaminatedwater|phylogeneticanalyses|sr|dem
Topic18	红树林等湿地生态预测分析	biodiversityconservation|n2oemission|mangrovewetland|ecologicalprocesses|stormsurge|n2ofluxes|trin|metalaccumulation
Topic19	湿地分类生态治理	mangroveforest|agriculturalwetland|landcoverchange|temporaldynamic|N20emission|coastalwetland|greatlake|scenario|soiltype
Topic20	自然和受控湿地的C、N循环模型的比较	combinedeffect|deltac13|denitrifier|localscale|differentwetlands|deltan15|foodresource|saltmarsh|sealevelrise|microbialdegradation
Topic21	湿地水质遥感评估	ecologicalrisk|satellitedata|dissimilatorynitratereduction|sourceidentification|coastalzone|tpremoval|Typha x glauca |situ measurement
Topic22	滨海湿地生态系统服务功能与管理	coastalecosystem|occupancy|horizontalsubsurfaceflow|co2|riverdelta|environmentalflow|polycyclicaromatichydrocarbonspah
Topic23	湿地社区生态学	environmentalgradients|restoredwetland|inhibition|ecologicalcondition|greywater|phytotoxicity|marshbird|typhadomingensis
Topic24	湿地生态修复	aquifers|porousmedia|cooccurrence|bacterialcommunitycomposition|wild|seedlingsurvival|leafareaindexlai|meteorologicaldata
Topic25	湿地生物对气候变化的反应	geographicallyisolatedwetlands|humanhealth|pca|stable-isotopes|climatewarming|species|n-addition|sodium|tree
Topic26	生物地球化学循环	soilorganiccarbon|waterhyacinth|phenotypicplasticity|biogeochemicalcycles|cd|coastalenvironment|growinginterest|nosZ-genes

4.3 新兴主题识别分析

4.3.1 一致连贯性分析

为了检测湿地领域主题的一致连贯性，通过时间划片并计算 2016-2020 年和 2011-2015 年的主题一致性指数，结果见图 6。主题一致连贯性指数的横坐标代表主题序号，纵坐标代表主题一致性指数计算结果。相邻时间区间内主题一致性指数均远高于设定阈值，说明利用主题模型确定的 26 个研究主题连接紧密，满足一致连贯性要求。

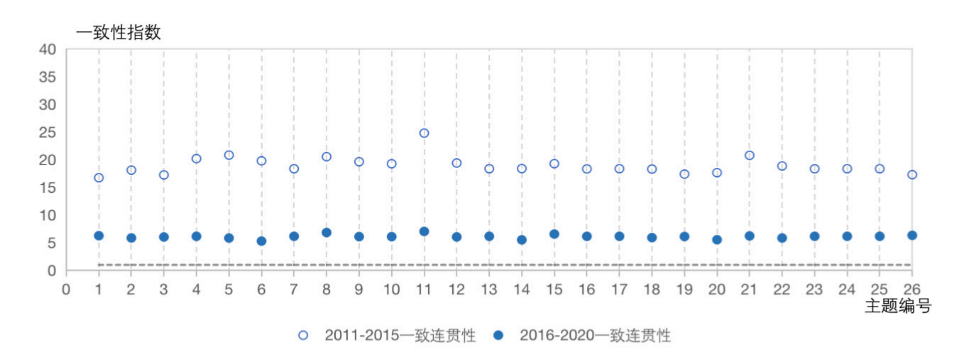


图 6 主题一致连贯性指数

ChinaXiv合作期刊
知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

4.3.2 潜在高成长力分析

针对潜在高成长力，笔者通过构建 ARIMA 模型来预测主题未来趋势。为避免潜在的自相关和异方差问题，预先对数据进行对数化处理，然后进行平稳性检验。检验类型上，分为趋势截距（c，t）、无趋势有截距（c，0）和无趋势无截距（0,0）3 种类型，根据显著性确定检验类型，平稳性检验结果见表 2。在主题强度序列中，进行差分处理后，主题 1、5、6、12、21 和 23 序列稳定，其余均为平稳序列，因此可以建模。

表 2 主体强度序列检验结果

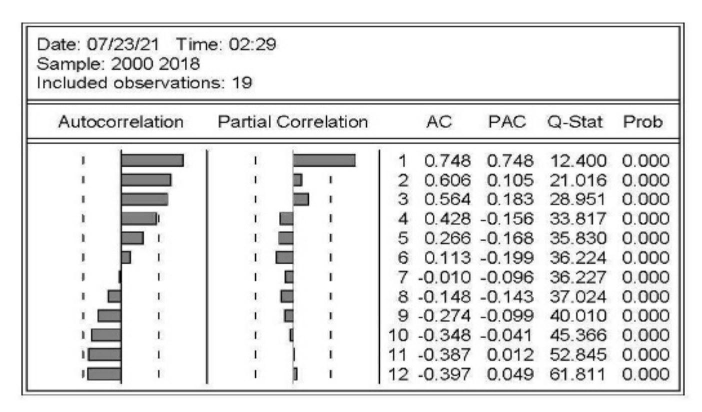
变量	检验类型	ADF统计值	5%统计值	P值	是否平稳
lnTopic1	（c,0,1）	-5.129 154	-3.710 482	0.004 0	截距平稳
lnTopic2	（c,t,0）	-6.472 344	-3.690 814	0.000 3	平稳
lnTopic3	（c,t,0）	-5.526 228	-3.690 814	0.001 7	平稳
lnTopic4	（c,t,0）	-3.322 204	-3.690 814	0.094 3	平稳
lnTopic5	（c,t,0）	-10.157 720	-3.052 169	0.000 0	截距平稳
lnTopic6	（c,t,1）	-3.564 050	-3.690 814	0.062 4	趋势截距平稳
lnTopic7	（c,t,0）	-3.502 356	-3.690 814	0.069 5	平稳
lnTopic8	（c,t,0）	-5.097 738	-3.690 814	0.003 8	平稳
lnTopic9	（c,t,0）	-3.7108 690	-3.690 814	0.048 2	平稳
lnTopic10	（c,t,0）	-4.091 747	-3.690 814	0.024 3	平稳
lnTopic11	（c,t,0）	-5.451 596	-3.690 814	0.002 0	平稳
lnTopic12	（c,0,1）	-10.268 850	-3.052 169	0.000 0	截距平稳
lnTopic13	（c,t,0）	-4.381 036	-3.690 814	0.014 3	平稳
lnTopic14	（c,t,0）	-3.714 111	-3.690 814	0.048 0	平稳
lnTopic15	（c,t,0）	-4.241 266	-3.690 814	0.018 4	平稳
lnTopic16	（c,t,0）	-3.542 071	-3.690 814	0.064 9	平稳
lnTopic17	（c,t,0）	-3.448 073	-3.690 814	0.076 3	平稳
lnTopic18	（c,t,0）	-4.975 529	-3.690 814	0.004 7	平稳
lnTopic19	（c,t,0）	-3.517 281	-3.690 814	0.067 7	平稳
lnTopic20	（c,t,0）	-4.485 202	-3.690 814	0.011 8	平稳
lnTopic21	（c,0,1）	-6.922 477	-3.052 169	0.000 0	截距平稳
lnTopic22	（c,t,0）	-3.675 896	-3.690 814	0.051 3	平稳
lnTopic23	（c,0,1）	-10.619 530	-3.052 169	0.000 0	截距平稳
lnTopic24	（c,t,0）	-4.346 055	-3.690 814	0.015 2	平稳
lnTopic25	（c,t,0）	-5.129 103	-3.690 814	0.003 6	平稳
lnTopic26	（c,t,0）	-11.538 490	-3.052 169	0.000 0	平稳

经过单位根检验，PCF 图、PACF 图定阶，结合信息准则（即 AIC、SC 和 HQ 最小个数最多原则）和参数比较，确定了 ARIMA 模型形式。由于建模期间过程数据较多，下面仅以表 3 展示最终模型参数定阶结果，并以主题 5 为例，展示建模流程。

表 3 ARIMA 时间序列模型搭建

主题	ACF图	PACF图	模型	主题	ACF图	PACF图	模型
主题1	拖尾	1阶截尾	ARIMA（1,0,0）	主题14	1阶截尾	1阶截尾	ARIMA（1,1,0）
主题2	3阶截尾	1阶截尾	ARIMA（1,0,0）	主题15	1阶截尾	1阶截尾	ARIMA（1,1,0）
主题3	拖尾	1阶截尾	ARIMA（1,0,0）	主题16	1阶截尾	1阶截尾	ARIMA（1,1,0）
主题4	拖尾	1阶截尾	ARIMA（1,1,0）	主题17	5阶截尾	1阶截尾	ARIMA（1,1,0）
主题5	3阶截尾	1阶截尾	ARIMA（1,1,0）	主题18	3阶截尾	1阶截尾	ARIMA（1,0,0）
主题6	5阶截尾	1阶截尾	ARIMA（1,1,0）	主题19	拖尾	1阶截尾	ARIMA（1,0,0）
主题7	6阶截尾	1阶截尾	ARIMA（1,0,0）	主题20	拖尾	1阶截尾	ARIMA（1,0,0）
主题8	拖尾	1阶截尾	ARIMA（1,1,0）	主题21	拖尾	1阶截尾	ARIMA（1,1,0）
主题9	1阶截尾	1阶截尾	ARIMA（1,0,0）	主题22	1阶截尾	1阶截尾	ARIMA（1,1,0）
主题10	拖尾	1阶截尾	ARIMA（1,0,0）	主题23	7阶截尾	1阶截尾	ARIMA（1,0,0）
主题11	拖尾	1阶截尾	ARIMA（1,0,0）	主题24	拖尾	1阶截尾	ARIMA（1,1,0）
主题12	拖尾	1阶截尾	ARIMA（1,1,0）	主题25	7阶截尾	1阶截尾	ARIMA（1,1,0）
主题13	7阶截尾	4阶截尾	ARIMA（1,1,0）	主题26	7阶截尾	8阶截尾	ARIMA（1,0,0）

如图 7 所示，主题 5 自相关图 3 阶截尾，偏自相关图 1 阶截尾，模型参数 p 应取 0-3 阶，参数 q 应取 0-1，可能存在 8 种可能的组合。通过信息准则比较，确定了模型的最优形式（见图 8）。据此展开主题趋势拟合和预测分析，图 9 左侧为基于 ARIMA 模型拟合的 2000-2018 年主题强度走势，呈现增长；右侧为 Topic5 未来 5 年主题走势预测结果，表现平稳。


Date: 07/23/21 Time: 02:29
Sample: 2000 2018
Included observations: 19

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.748	0.748	12.400	0.000
		2	0.606	0.105	21.016	0.000
		3	0.564	0.183	28.951	0.000
		4	0.428	-0.156	33.817	0.000
		5	0.266	-0.168	35.830	0.000
		6	0.113	-0.199	36.224	0.000
		7	-0.010	-0.096	36.227	0.000
		8	-0.148	-0.143	37.024	0.000
		9	-0.274	-0.099	40.010	0.000
		10	-0.348	-0.041	45.366	0.000
		11	-0.387	0.012	52.845	0.000
		12	-0.397	0.049	61.811	0.000

图 7 主题 5 建模 PAC 和 PACF 图

Dependent Variable: LNTOPIC2
Method: Least Squares
Date: 07/23/21 Time: 03:17
Sample (adjusted): 2001 2018
Included observations: 18 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.354134	0.244236	1.449970	0.1664
LNTOPIC2(-1)	0.771580	0.175189	4.404265	0.0004

R-squared	0.547991	Mean dependent var	1.371429
Adjusted R-squared	0.519741	S.D. dependent var	0.485903
S.E. of regression	0.336734	Akaike info criterion	0.765394
Sum squared resid	1.814240	Schwarz criterion	0.864325
Log likelihood	-4.888549	Hannan-Quinn criter.	0.779036
F-statistic	19.39755	Durbin-Watson stat	2.233050
Prob(F-statistic)	0.000443		

图 8 主题 5 模型信息准则及参数

知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

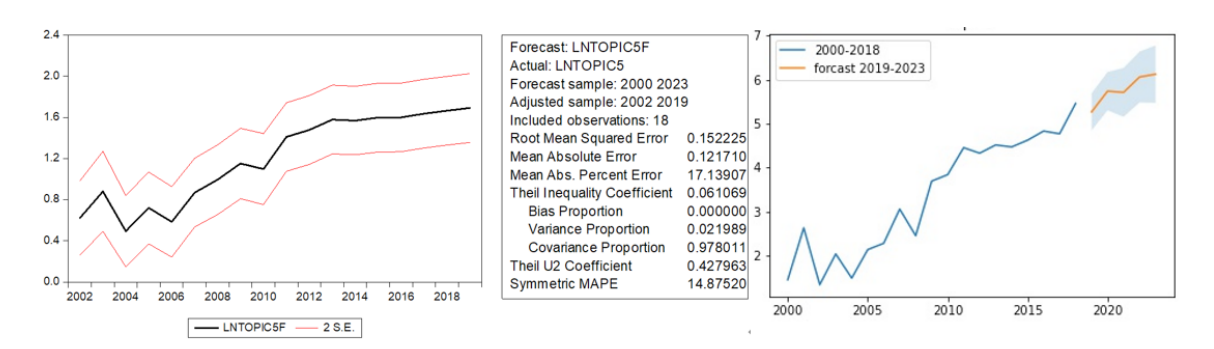
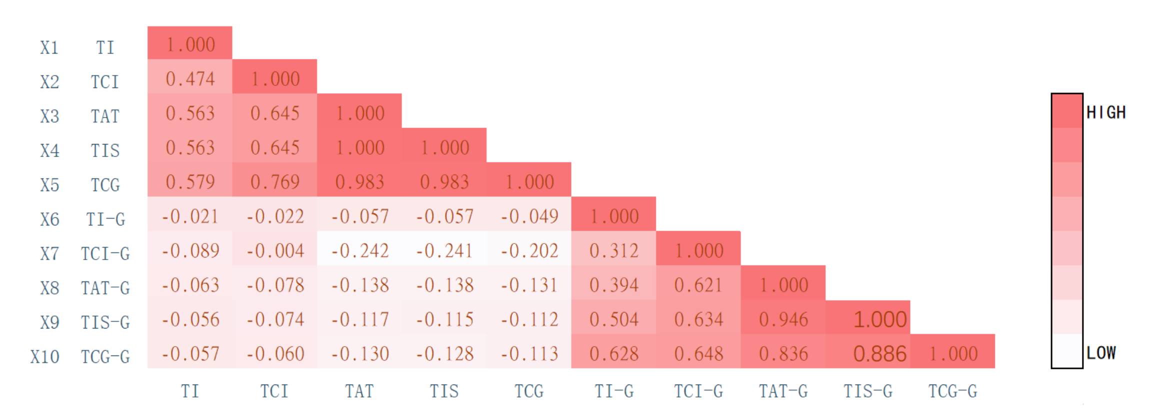


图 9 主题 5 基于 ARIMA 的趋势拟合预测

4.3.3 影响力和增长性分析

时序全局主成分分析利用综合变量来取代原有的全局变量，能抓住主要影响特征 [38]。通过计算 2001-2018 各年度度量指标，得到 260 × 18 的时序数据表，共 4 680 条数据，指标间存在相关性（见图 10）。为消除量纲的影响，采取标准化处理，巴特利球度检验统计量为 9 135.283，p 值接近 0，KMO 检验值大于 0.7，适合主成分分析。



		TI	TCI	TAT	TIS	TCG	TI-G	TCI-G	TAT-G	TIS-G	TCG-G
X1	TI	1.000									
X2	TCI	0.474	1.000								
X3	TAT	0.563	0.645	1.000							
X4	TIS	0.563	0.645	1.000	1.000						
X5	TCG	0.579	0.769	0.983	0.983	1.000					
X6	TI-G	-0.021	-0.022	-0.057	-0.057	-0.049	1.000				
X7	TCI-G	-0.089	-0.004	-0.242	-0.241	-0.202	0.312	1.000			
X8	TAT-G	-0.063	-0.078	-0.138	-0.138	-0.131	0.394	0.621	1.000		
X9	TIS-G	-0.056	-0.074	-0.117	-0.115	-0.112	0.504	0.634	0.946	1.000	
X10	TCG-G	-0.057	-0.060	-0.130	-0.128	-0.113	0.628	0.648	0.836	0.886	1.000

图 10 影响力和增长性各成分相关性分析

计算全局主成分分析的初始解及因子解情况，依据特征值大于 1 的原则，选取主成分 F1 和 F2，二者分别携带 43.375% 和 32.519% 的原始数据信息。第一主成分中 5 项影响力指标均为正值且有较大的载荷，构成影响力综合因子。第二主成分更多地反映了主题增长性情况，构成增长性因子。

利用成分得分系数得到两类主成分的解析表达式，如下所示：

F1=0.130X1+0.146X2+0.191X3+0.191X4+0.193X5-0.087X6-0.130X7-0.138X8-0.139X9-0.139X10 公式（9）

F2=0.124X1+0.146X2+0.152X3+0.153X4+0.162X5+0.159X6+0.157X7+0.214X8+0.227X9+0.226X10 公式（10）

根据累计贡献度归一化处理，计算综合评价指标权重，可以进一步得到新兴主题影响力和增长性综合评价的表达式，如下所示：

F=0.1272X1+0.1461X2+0.1744X3+0.1744X4+0.1795X5+0.0186X6-0.0069X7+0.0127X8+0.0177X9+0.0171X10 公式（11）

为了更好地解释主成分的现实意义，可以通过数据标准化和各主成分得分计算观察主题

二维分布情况，如图 11 所示。主题 7、13、16、23、24、25 和 26 等呈现出高增长与高影响的协同发展效应，表现高增长新兴主题可以取得更多的科学影响力；主题 1、2、4、5、6、10、11、12、14、18、19 和 21 等，其增长性和影响力呈现出一定的替代效应；主题 3、8、9、12 和 15 等分布接近原点，影响力和增长性特征发展较为稳定。

注：横轴为影响力维度，纵轴为增长性维度

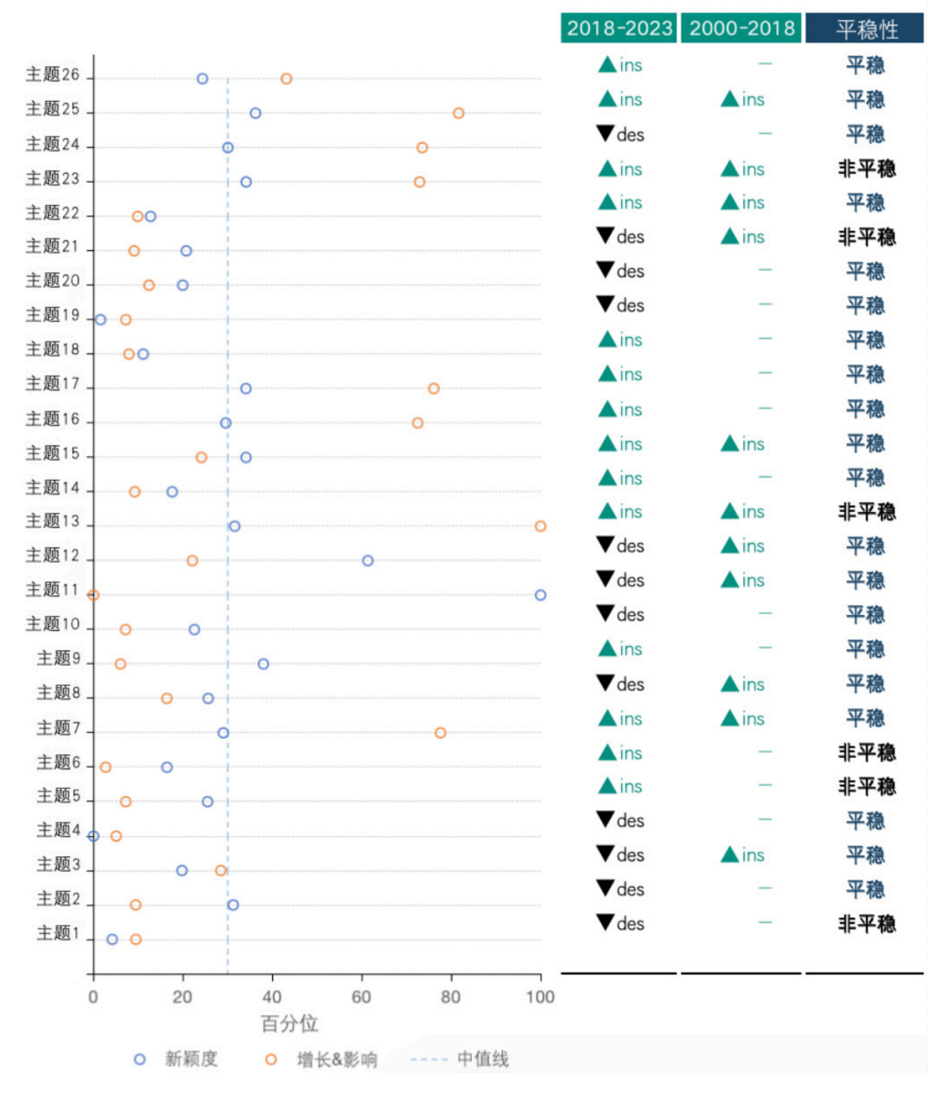
图 11 湿地领域主题增长性及影响力因子动态分布

4.3.4 新兴主题识别结果

综合湿地领域主题各维度的特征，可以发现：①通过主题模型计算得到的 26 个主题均满足一致连贯性特征的要求。②潜在高成长性分析结果显示，在 2000-2018 年里，主题强度大部分呈现平稳或上升的态势；在未来 5 年中，主题 5、6、7、9、13、14、15、16、17、18、22、23、25 和 26 拥有显著的潜在高成长力，预计发展态势向好。③新颖度方面表现良好的主题包括主题 2、7、9、11、12、13、15、16、17、23 和 25。④联合分析增长性和影响力，主题 3、7、13、16、17、23、24、25 和 26 拥有具有较好的特征表现。

新兴主题多维识别结果如图 12 所示，结果表明，在湿地领域符合新兴主题定义的主题为主题 7、13、15、16、17 和 25，即退化湿地系统恢复、湿地微生物基因研究、湿地物质平衡 / 湿地生态补偿、湿地定量勘查研究、湿地菌群系统治理分析和湿地生态对气候变化响应分析。

知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020



	2018-2023	2000-2018	平稳性
主题26	▲ins	—	平稳
主题25	▲ins	▲ins	平稳
主题24	▼des	—	平稳
主题23	▲ins	▲ins	非平稳
主题22	▲ins	▲ins	平稳
主题21	▼des	▲ins	非平稳
主题20	▼des	—	平稳
主题19	▼des	—	平稳
主题18	▲ins	—	平稳
主题17	▲ins	—	平稳
主题16	▲ins	—	平稳
主题15	▲ins	▲ins	平稳
主题14	▲ins	—	平稳
主题13	▲ins	▲ins	非平稳
主题12	▼des	▲ins	平稳
主题11	▼des	▲ins	平稳
主题10	▼des	—	平稳
主题9	▲ins	—	平稳
主题8	▼des	▲ins	平稳
主题7	▲ins	▲ins	平稳
主题6	▲ins	—	非平稳
主题5	▲ins	—	非平稳
主题4	▼des	—	平稳
主题3	▼des	▲ins	平稳
主题2	▼des	—	平稳
主题1	▼des	—	非平稳

图 12 新兴主题特征维度分析

5 新兴主题特征关联分析

新兴主题具有发展成为未来热点主题的趋势 [39]，是前沿主题孵化的基床 [40]。在新兴主题识别的基础上，深度挖掘新兴主题关联特征的长期关系，可以更好地认识新兴主题，具有一定的现实意义。

本文立足于科学评价体系，选取能反映研究强度、研究绩效、研究机构多样化程度和成果传播能力的主要计量指标，针对新兴主题组成的面板数据，进行主题特征（包括引文特征、作者特征、机构规模以及学科丰富性等）关联分析。为规避可能存在的异方差情况，对数据进行对数化处理后，完成LLC平稳性检验，其中，主题多学科特征存在单位根过程，即一阶单整，其余变量均为零阶单整。

5.1 长期均衡分析：协整分析

由于采用的数据并非同阶单整，需要经协整检验确定长期稳定关系。在 Kao-test 协整检验中，原假设为主题强度与主题特征数据不存在协整关系。根据 DF 和调整的 ADF 等 5 个检验统计量的显著性比较，结论均为拒绝原假

设（见表 4），即存在协整关系，主题强度序列与主题各维度外部特征之间存在长期稳定关系，可以对观测变量进行进一步的因果关系分析诊断。

表 4 Kao-test 协整检验

Kao test for cointegration			
Ho:No cointegration		Number of panels=	6
Ha:All panels are cointegrated		Number of periods=	19
Cointegrating vector:	Same		
Panel means:	Included	Kernel:	Bartlett
Time trend:	Not included	Lags:	1.83(Newey-West)
AR parameter:	Same	Augmented lags:	
		Statistic	p-value
Modified Dickey-Fuller t		-3.610 2	0.000 2
Dickey-Fuller t		-5.250 7	0.000 0
Augmented Dickey-Fuller t		-7.705 2	0.000 0
Unadjusted modified Dickey-Fuller t		-5.700 6	0.000 0
Unadjusted Dickey-Fuller t		-5.822 1	0.000 0

根据协整方程可得：主题机构的增长、主题作者数量的增长和主题被引频次的增加，与主题强度在长呈现正向的均衡关系；主题学科丰富性的增加与主题强度在长期呈现负向的均衡关系，如表 5 所示：

表 5 协整方程

Cointegrating Equation(s):			Log likelihood	964.444 2
Normalized cointegrating coefficients (standard error in parentheses)				
LNTI	LNTCI	LNTCG	LNTAT	LNTIS
1.000 000	234.143 1	-920.788 2	469.822 8	227.383 8
	(32.655 9)	(129.873)	(208.967)	(276.913)
Adjustment coefficients (standard error in parentheses)				
D(LNTI)	0.001 462			
	(0.000 23)			
D(LNTCI)	0.006 427			
	(0.000 69)			
D(LNTCG)	0.001 084			
	(0.000 24)			
D(LNTAT)	0.001 080			
	(0.000 23)			
D(LNTIS)	0.001 183			
	(0.000 24)			

知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

5.2 格兰杰因果关系检验

格兰杰因果检验是一种预测能力的检验，其基本原理为：假设变量甲和乙间存在互相影响，如果甲的滞后期变量对乙有显著影响，那么甲是乙的格兰杰原因，反之则反[41]。确认主题强度与各维度特征之间存在协整关系后，由于作用方向不明，笔者首先利用Pvar模型确定最优滞后阶数为3，接着采用A. Juodis 等[42]提出的格兰杰因果检验方法对变量的外生性进行检验，确定主题各维度特征对主题强度是否拥有解释能力，若无则需调整。

原假设为被解释变量主题联合维度特征对主题强度无显著性影响，检验结果见表6，机构、作者、引用和学科丰富性对主题强度的三阶滞后项对主题强度的影响显著性水平有所差异，但其联合作用的临界值小于0.05，说明4个变量的联合变化是主题强度变化的内生因素。为研究变量之间具体的因果关系，笔者进行进一步的格兰杰因果检验，结果见表7。

表6 格兰杰外生性检验

Juodis, Karavias and Sarafidis(2021) Granger non-causality test results:						
Number of units=	6			T=	18	
Number of lags =	3			BIC=	342.544 2	
HPJ Wald test：	76 510.84		pvalue_HPJ：		0.000 0	
H0:	Selected covariates do not Granger-cause ti.					
H1:	H0 is violated.					
Results for the Half-Panel Jackknife estimator						
	Coef.	Std.Err.	z	P>|z|	[95% Conf. Interval]	
tci						
L1.	-0.032 670	0.003 953	-8.27	0.000	-0.040 420	-0.024 930
L2.	-0.026 340	0.003 456	-7.62	0.000	-0.033 110	-0.019 570
L3.	-0.146 610	0.003 222	-45.50	0.000	-0.152 920	-0.140 290
tis						
L1.	2.650 179	0.125 868	21.06	0.000	2.403 481	2.896 877
L2.	-6.499 070	0.142 101	-45.74	0.000	-6.777 580	-6.220 550
L3.	15.612 260	0.168 684	92.55	0.000	15.281 640	15.942 870
tat						
L1.	-2.534 910	0.058 553	-43.29	0.000	-2.649 670	-2.420 150
L2.	1.831 435	0.063 434	28.87	0.000	1.707 107	1.955 762
L3.	-5.879 390	0.071 325	-82.43	0.000	-6.019 180	-5.739 590
tcg						
L1.	3.946 251	0.092 628	42.60	0.000	3.764 703	4.127 798
L2.	3.033 879	0.073 701	41.16	0.000	2.889 429	3.178 330
L3.	-4.039 500	0.066 424	-60.81	0.000	-4.169 690	-3.909 310

表 7 Granger 因果关系检验结果

零假设	观测量	F 统计量	P 值	结论
lNTI 不是 LITIS 的 Granger 原因	114	3.340 00	0.070 3	拒绝
lNTI 不是 LITCI 的 Granger 原因	114	3.488 47	0.063 9	拒绝
lNTI 不是 LITAT 的 Granger 原因	114	2.975 39	0.089 2	拒绝
lNTI 不是 LITCG 的 Granger 原因	114	1.018 18	0.390 4	接受
lNTIS 不是 LITI 的 Granger 原因	114	8.040 20	0.006 1	拒绝
lNTCI 不是 LITI 的 Granger 原因	114	2.576 47	0.117 2	接受
lNTAT 不是 LITI 的 Granger 原因	114	3.765 67	0.053 8	拒绝
lNTCG 不是 LITI 的 Granger 原因	114	3.090 81	0.082 7	拒绝

分析表 7Granger 因果关系检验结果，可得出如下结论：

（1）对于湿地领域的新兴主题而言，主题强度和主题机构数量、主题作者数量之间存在双向的格兰杰因果关系。这说明，领域内研究学者的增长促进了领域新兴主题的发展，主题强度的增长也吸引了新的一批学者展开相关的研究，结果验证了集群效应，说明人才发展与主题发展属于相辅相成的主动模式。这从侧面反映出湿地领域相关研究支持机构制定研究激励政策的有效性，在未来发展学科主题时应考虑项目为先、人才为本的执行思路。

（2）在湿地领域中，主题强度和主题学科丰富性数量、主题引用间存在单向的因果关系，即主题强度的良好发展是主题学科丰富性的原因，但学科丰富性不是主题强度良好发展的原因；主题强度增长是主题引用频次增加的原因，而主题引用频次增加是主题强度变化的原因。其现实含义为，主题强度对主题丰富性有着单方面作用，主题强度随着时间发展而不断扩张，促进了湿地领域学科的多元化发展；然而，湿地领域学科丰富性的发展并没有明显优化主题强度的增长，这说明，通过促进学科丰富性的增加并不能够直接地促进该领域主题强度的良性发展，在湿地领域内盲目追求学科丰富性，可能导致主题分散化较为严重，难以做到“大而精”。此外，引用情况在一定程度上代表着主题关注度的转移，主题强度增长对于引用的拉动作用在短期内因果关系不显著，反观主题引用频次对主题强度发展的影响，可以发现，引用频次增加对主题强度发展的促进效果显著，是该领域主题强度发展的“风向标”。

6 讨论

从论文数据中，笔者提出了一套基于新兴主题特征的识别与关联分析方法。在特征提取方面，结合新兴主题相关理论与实践，在新颖性等方面做出了改良，加入潜在高成长性指标，并针对影响力和增长性选取了较为全面的特征考量方案。本研究通过主题模型提取研究主题与主题分布，采用趋势预测模型与分析方法分析主题未来趋势，结合全局主成成分析刻画主题增长性和影响力动态发展路径，根据主题综合表现情况完成新兴主题的识别。为更好地识别新兴主题，笔者利用协整分析和格兰杰因果检验，针对新兴主题的特征关联关系进行挖掘，研究发现，主题强度与机构数量、作者规模间存在双向的关联效应，主题引用频次对主题发展存在正向的影响，主题强度对主题多样性产生单向的促进作用，由此，笔者提出应坚持项目为先、人才为本的创新政策执行思路，以及关于如何发展新兴主题的一些思考。笔者在特征科学性和识别全面性上进行了反复考量，综合选用自然语言处理、多元统计分析和时间序

知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

列分析方法，确定了新兴主题识别与特征分析方法，该方法对于客观认识领域内研究主题动态、展开科研布局决策等具有一定的参考价值。

笔者提出的新兴主题识别分析方法主要从科学文献角度展开，由于新兴主题是一个领域内研究内容的全面特征，其研究价值体现在科技、政策和经济等各个方面，而文献只是反映研究主题创新变化的一个重要对象，除科学文献外，还包括政策文本和专利数据等研究对象。因此，未来研究可以尝试将多源文本融合进行综合的新兴主题识别研究。

参考文献：

[1] 刘自强 , 王效岳 , 白如江 . 多维度视角下学科主题演化可视化分析方法研究——以我国图书情报领域大数据研究为例 [J]. 中国图书馆学报 , 2016, 42(6): 67-84.

[2] 王山 . 研究前沿探测方法进展 [J]. 情报科学 , 2019, 37(10): 164-169.

[3] XU H, WINNINK J, YUE Z, et al. Multidimensional scientometric indicators for the detection of emerging research topics[J]. Technological forecasting and social change, 2021, 163: 1-25.

[4] LU C, HOU H, DING Y, et al. Review of internatinonal studies on discovering emerging topics[J/OL]. Journal of the China Society for Scientific and Technical Information, 2019[2021-09-13]. http: //en.cnki.com.cn/Article_en/CJFDTotal-QBXB201901011.htm.

[5] LIU G Y, HU J M, WANG H L. A co-word analysis of digital library field in China[J]. Scientometrics, 2012, 91(1): 203-217.

[6] CHI R, YOUNG J. The interdisciplinary structure of research on intercultural relations: a co-citation network analysis study[J]. Scientometrics, 2013, 96(1): 147-171.

[7] SONG M, KIM S Y. Detecting the knowledge structure of bioinformatics by mining full-text collections[J]. Scientometrics, 2013, 96(1): 183-201.

[8] 钟辉新 . 新兴趋势探测研究综述 [J]. 现代情报 , 2017, 37(12): 162-167.

[9] XU S, HAO L, AN X, et al. Emerging research topics detection with multiple machine learning models[J]. Journal of informetrics, 2019, 13(4): 100983.

[10] 刘小玲 , 谭宗颖 . 新兴技术主题识别方法研究进展 [J]. 图书情报工作 , 2020, 64(11): 145-152.

[11] DE SOLLA PRICE D J. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.

[12] OHNIWA R L, HIBINO A, TAKEYASU K. Trends in research foci in life science fields over the last 30 years monitored by emerging topics[J]. Scientometrics, 2010, 85(1): 111-127.

[13] TU Y N, SENG J L. Indices of novelty for emerging topic detection[J]. Information processing & management, 2012, 48(2): 303-325.

[14] ROTOLO D, HICKS D, MARTIN B R. What is an emerging technology?[J]. Research policy, 2015, 44(10): 1827-1843.

[15] WANG Q. A bibliometric model for identifying emerging research topics[J]. Journal of the Association for Information Science and Technology, 2018, 69(2): 290-304.

[16] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents[J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.

[17] CHEN C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377

[18] 王燕鹏 . 国内基于主题模型的科技文献主题发现及演化研究进展 [J]. 图书情报工作 , 2016, 60(3): 130-137.

[19] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(4/5): 993-1022.

[20] GERRISH S, BLEI D M. A language-based approach to measuring scholarly impact[C/OL]. [2021-08-10]. https: //openreview.net/forum?id=HJ-9EsbdWr.

[21] XU M, LI G, WANG X.Detecting emerging topics by exploiting probability burst and association rule mining: a case study of library and information science[J]. Malaysian journal of library & information science, 2020, 25(1): 47-66.

[22] 李静 , 徐路路 . 基于机器学习算法的研究热点趋势预测模型对比与分析——BP 神经网络、支持向量机与 LSTM 模型 [J]. 现代情报 , 2019, 39(4): 23-33.

[23] 白敬毅 , 颜端武 , 陈琼 . 基于主题模型和曲线拟合的新兴主题趋势预测研究 [J]. 情报理论与实践 , 2020, 43(7): 130-136, 193.

[24] KONTOSTATHIS A, GALITSKY L M, POTTENGER W M, et al. A survey of emerging trend detection in textual data mining[C]//BERRY M W. Survey of text

mining: clustering, classification, and retrieval. New York: Springer, 2004: 185-224.

[25] LEE C, KWON O, KIM M, et al. Early identification of emerging technologies: a machine learning approach using multiple patent indicators[J]. Technological forecasting and social change, 2018, 127: 291-303.

[26] 岳丽欣 , 周晓英 , 陈旖旎 . 基于 ARIMA 模型的信息构建研究主题趋势预测研究 [J]. 图书情报知识 , 2019(5): 54-63, 72.

[27] 岳丽欣 , 刘自强 , 胡正银 . 面向趋势预测的热点主题演化分析方法研究 [J]. 数据分析与知识发现 , 2020, 4(6): 22-34.

[28] 刘自强 , 许海云 , 岳丽欣 , 等 . 面向研究前沿预测的主题扩散演化滞后效应研究 [J]. 情报学报 , 2018, 37(10): 979-988.

[29] 白如江 , 刘博文 , 冷伏海 . 基于多维指标的未来新兴科学研究前沿识别研究 [J]. 情报学报 , 2020, 39(7): 747-760.

[30] 王茜 , 谭宗颖 , 钱力 . 科学研究社会影响力评价综述 [J]. 图书情报工作 , 2015, 59(14): 143-148.

[31] GONZALEZ-ALCAIDE G, GORRAIZ J, HERVAS-OLIVER J L. On the use of bibliometric indicators for the analysis of emerging topics and their evolution: spin-offs as a case study[J]. Profesional de la informacion, 2018, 27(3): 493-510.

[32] GUO H, WEINGART S, BÖRNER K. Mixed-indicators model for identifying emerging research areas[J]. Scientometrics, 2011, 89(1): 421-435.

[33] 万伦来 , 干俊峰 , 余晓钰 . 基于 Matlab 的时序全局主成分分析方法及应用 [J]. 华东经济管理 , 2010, 24(1): 150-153.

[34] CHEN B, TSUTSUI S, DING Y, et al.Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval[J]. Journal of informetrics, 2017, 11(4): 1175-1189.

[35] ENGLE R F, GRANGER C W J. Co-integration and error correction: representation, estimation, and testing[J]. Econometrica, 1987, 55(2): 251.

[36] KAO C W, CHIANG M H. On the estimation and inference of a cointegrated regression in panel data[J]. Advances ecoometrics, 2000, 15: 179-222.

[37] DUMITRESCU E I, HURLIN C. Testing for granger non-causality in heterogeneous panels[J]. Economic modelling, 2012, 29(4): 1450-1460.

[38] 乔峰 , 姚俭 . 时序全局主成分分析在经济发展动态描绘中的应用 [J]. 数理统计与管理 , 2003(2): 1-5.

[39] 罗瑞 , 许海云 , 董坤 . 领域前沿识别方法综述 [J]. 图书情报工作 , 2018, 62(23): 119-131.

[40] 黄晓斌 , 吴高 . 学科领域研究前沿探测方法研究述评 [J]. 情报学报 , 2019, 38(8): 872-880.

[41] 谷祖莎 . 我国贸易开放与二氧化碳排放的关系研究 [J]. 学术论坛 , 2012, 35(8): 109-112.

[42] JUODIS A, KARAVIAS Y, SARAFIDIS V. A homogeneous approach to testing for Granger non-causality in heterogeneous panels[J]. Empirical Economics, 2021, 60(1). DOI: 10.1007/s00181-020-01970-9.

作者贡献说明：
李雅倩：研究框架搭建，数据分析，文章撰写；
孙玉玲：论文指导，成稿修改；
赵婉雨：数据收集与预处理。

ChinaXiv合作期刊
知识管理论坛，2022（3）:228-247
DOI: 10.13266/j.issn.2095-5472.2022.020

Research on Emerging Topic Recognition and Feature Association Based on Topic Model and Time Series Analysis

Li Yaqian[1,2] Sun Yuling[1,2] Zhao Wanyu[1]
[1]National Science Library, Chinese Academy of Sciences, Beijing 100080
[2]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100080

Abstract: [Purpose/Significance] Carrying out research on emerging research topics(ERT) identification and scientifically and effectively discovering their characteristic correlation laws can better serve practical needs and give play to the innovative supporting role of sci-tech information research on the development of disciplines. Aiming at discovering emerging research topic(ERT) and its characteristic correlation effect scientifically and effectively, this paper carries out ERT identification and feature analysis, while realizing the innovative supporting role of sci-tech information work. **[Method/Process]** Starting from the definition of the features of ERT, this paper established the methodological framework of ERT identification by using natural language processing, global principal component analysis and time series analysis. Based on the relevant theories and practices of emerging topic identification and scientific impact assessment, this thesis quantified the characteristics of the topic’s consistency, novelty, influence, and growth. On the basis of emerging themes identification, the law of the development of emerging themes in the target field is deeply excavated. Granger causality test and cointegration analysis were used to explore the long term equilibrium and the correlation effects of their characteristics. **[Result/Conclusion]** This paper proposes a method to identify ERT and their correlation feature analysis. In order to verify the effectiveness and feasibility of this method, the field of wetland was selected to carry out empirical research. Combined with the topic identification and feature correlation effect analysis, the final result depicted the dynamic development path of subject science influence in this field, while putting forward some advices on developing emerging topics from the perspective of associated characteristics.

Keywords: trend forecasting　emerging research topic identification　characteristic correlation effect　cointegration analysis　panel data analysis